Applied Computing and Informatics 15 (2019) 182–190

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

Predicting rank for scientific research papers using supervised learning

Mohamed El Mohadab *, Belaid Bouikhalene, Said Safi

Department of Mathematics and Informatics, Laboratory of Innovation in Mathematics and Application and Information Technologies (LIMATI), Polydisciplinary Faculty, Sultan Moulay Slimane University, PB 592 Beni Mellal, Morocco

ARTICLE INFO

Article history:
Received 10 September 2017
Revised 14 February 2018
Accepted 17 February 2018
Available online 6 March 2018

Keywords:
Scientific research
Ranking scientific research papers
Data mining
Supervised learning
Multilayer perceptron algorithm

ABSTRACT

Automatic data processing represents the future for the development of any system, especially in scientific research. In this paper, we describe one of the automatic classification methods applied to scientific research as a supervised learning task. Throughout the process, we identify the main features that are used as keys to play a significant role in terms of predicting the new rank under the supervised learning setup. First, we propose an overview of the work that has been realized in ranking scientific research papers. Second, we evaluate and compare some of state-of-the-art for the classification by supervised learning, semi-supervised learning and non-supervised learning. During the preliminary tests, we have obtained good results for performance on realistic corpus then we have compared performance metrics, such as NDCG, MAP, GMAP, F-Measure, Precision and Recall in order to define the influential features in our work.

1. Introduction

Due to the fast development of information and communications technologies, the university has firmly decided to facilitate the access and treatment for all processes, especially in scientific research in order to assist PhD students, professors and administrative staff to deal with digital services that they need.

In recent years, research in ranking scientific research papers (SRP) from diversified fields of research has become a very important task because of the exponential growing of daily publication in journals and conferences, exceeding 50 million papers. Also, predicting the future of any system represents another challenge that we can face generally, but mixing both problems is the case that we address in this research by predicting the new rank of scientific research paper.

Using machine learning [1] in ranking scientific research papers is a crucial research direction, because it contains distinct classes of supervised learning algorithms [2] with regard to prediction. Between the main important algorithms used in linear classifiers, we choose to work with Multilayer Perceptron Algorithm [3], SMO Classifier [4], and Kstar Classifier [5]. The three classifiers represent the neuron network with nodes and edges as papers and citations between the different authors in a similar set of information. The major reason for this choice is that a research paper network is a concrete example of the relation where the researcher collaborates with other research communities in the scientific domains in order to achieve their goals.

The related work gives us a vision on the approaches and methods for the classification of scientific research papers, and which is grouped into two major axes: the first axis is ranking according to the query and the second is ranking according to the technical analysis link. The limitation of these two main axes classifies the existing papers to us but it does not propose a contribution concerning the future classification. The novelty brought in this work is manifested through the prediction of the future classification being based on the existing papers that will offer the researcher the paper with the highest rank in this field.

The rest of this paper is organized as follows: in the next section we review the related work in ranking scientific research papers. Section 3 shows the state of the art for the learning methods. Section 4 describes Methods. Section 5 shows Results and discussion. Finally, in section 6, we conclude and describe future research directions.

2. Related work

In the last few years, there has been a growing interest in ranking scientific research papers as one of the pillars of research in

* Corresponding author.
E-mail address: m.elmohadab@gmail.com (M. El Mohadab).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.aci.2018.02.002
2210-8327/

ranking in general. Among the most used tools to rank journals, we find the impact factor [6], which is an approximation of the average number of citations within a year given to the set of papers belonging to a journal published in the two preceding years.

Previous studies in ranking indicate that we have two axes; the first axis represents the relevance ranking algorithm which matches the ranking according to the query. Among the studies we find:

- Vector Space Model [7]: is the way of representing text documents as vectors of identifiers. It is used in information retrieval, information filtering, and relevancy and indexing ranking.
- Latent semantic analysis [8]: is the way of analyzing the relation between a set of concepts related to the term and a set of documents. This technique is used in natural language processing.
- Okapi BM25 [9]: is the way of matching documents by a ranking function in search engines according to their relevance with a search query. It is totally based on the probabilistic retrieval framework.
- Boolean Ranking Model [10]: is the way of searching the user's query in the existent set based on classical set theory and Boolean logic.

The second axis represents an important ranking model which ranks according to the link analysis technique. Among the models we find:

- HITS [11]: is an analysis algorithm, the basic idea is that a web page serves two purposes: to provide information and or to suggest links to pages on a topic.
- PageRank [12]: is an algorithm working by calculating the number of links to a page and also their quality to determine a close estimation of the importance of the relevant documents.

Most algorithms used to rank SRP are divergent from PageRank or HITS, we find:

- Topic Rank [13]: clusters papers into topics; between the main factors used, we have: topic, citation, date of publication, title, and keyword.
- Cite Rank [14]: is an algorithm working by ranking citation networks based on their topology, between the main factors used, we have: citation, title, and date of publication.
- PTRA [15]: gives the paper age a higher impact, and depends highly on time of publication to rank the papers; among the main factors used, we have: citation, publication venue, and publication date.

However, all researchers have concluded that both types of ranking algorithms (the relevant/the important) have some limitations, especially the relevance algorithm that is not used any more in ranking algorithms. However, studies on predicting a new rank are still lacking. In the next section, we will review the state of the art for the learning methods.

3. State of the art for the learning methods

In this section, we will provide a sort of overview of what has been found in the literature concerning the models which we have been using in our work. There are three families of differential learning methods: supervised learning, which requires prior labeling of class data so that the model can train on them; unsupervised learning (clustering), without a prior information input; and semi-supervised learning mode, which jointly manipulates unlabeled and labeled data.

3.1. Supervised approach

Supervised Approach is an automatic learning technique which leads automatically to produce rules from a learning database containing examples of cases already dealt with. Therefore, its aim is to generalize for unknown inputs that it has been apt to learn from the data which are already handled by experts; the purpose is to use this to determine a compact representation of the function of prediction, which at a new input x associates an output S (x). The three main approaches related to supervised learning are:

- Neural Networks [16].
- Hidden Markov Model [17].
- Support Vector Machines [18].

The Neural Networks [16] is generally defined by three types of parameters:

- The interconnection pattern.
- The activation function.
- The learning process.

The Hidden Markov Model [17] is defined by two stochastic processes: a Markov chain is defined by a set of states and the transitions between the different states, so-called emission probabilities connected with each state. We will bring into focus the decision-making process, which is described by:

- A finite set S of discrete states denoted s.
- A finite set A of actions denoted a.
- A transition function P: $S \times A \rightarrow P(S)$ where P (S) is the set of distributions of probability on S.

The Support Vector Machine [18] offers, in particular, a good approximation of the fundamental of minimization of structural risk. The method depends on the following ideas:

- The data is projected in a large space by a transformation based on a linear, polynomial or Gaussian kernel.
- The classes are disconnected by linear classifiers that maximize the margin in the transformed space.
- The hyper planes can be determined by means of a few points which will be called "support vectors".

The Boosting [19] is summarized as follows:

- A large set of simple features.
- Initialization weights for training sets.
- For T rounds:
  - Normalize the weights.
  - For features from the set, train a classifier with a single feature and examine the training error.
  - Determine the classifier with the lowest error.
  - Update the weights of the training sets.
- The final classifier is the linear combination of the T classifiers.

3.2. Unsupervised approach

There are different reasons for choosing this type of learning such as the charge of developing manual labeling and the search for discriminatory characteristics in the first study or characteristics which grow over time. Unsupervised learning [20] is often

treated as a density estimation problem; the two main approaches used in unsupervised learning are:

- K-Means Clustering [21].
- Fuzzy C-Means [22].

The k-means clustering [21] is summarized as follows:

- Place K points into the space expressed by the objects clustered;
- Assign each object to the near centroid;
- Recalculate the locations of the K centroids;
- Repeat until the centroids are fixed. The metric is then calculated.

The fuzzy c-means algorithm [22] is very identical to the k-means algorithm:

- Appoint a number of clusters.
- Each point is given a random coefficient.
- Repeat the algorithm until convergence.

3.3. Semi-supervised approach

To perform generic tasks of supervised learning while exploiting some labeled data simultaneously with multiple raw data. The first idea is using a non-supervised context of the outputs predicted by the system itself in order to construct the desired outputs by applying a supervised technique. This approach is known as the directed decision. The second idea depends on the simultaneous use of two classifiers. They alternately act as a teacher and a pupil in an algorithm, iterative learning: the output calculated by one will be taken as the appropriate output by the other and reciprocally until convergence.

The learning criterion is here to optimize the coherence between the two classifiers. This approach is known as self-supervision. The two main approaches to semi-supervised learning [23] are:

- Co-Training [2]: It is a machine learning algorithm used when there are only some labeled data and large amounts of unlabeled data. One of its uses is in text mining for search engines.
- Co-Boosting [24]: It may be seen as a combination of co-training and boosting.

Moreover, we have compared distinct machine learning algorithms. In our case of study, we have chosen to work with the supervised learning approach especially with the neuron network represented by the Multilayer Perceptron classifier [3]; the reasons for this choice are:

- Presentation of a drive to the network.
- Comparison of the network output with the targeted output.
- Calculation of the error at the output of each neuron belonging to the network.
- Definition of the increase or decrease required to obtain this value.
- Adjustment of the weight of each connection to the lowest local error.
- Granting blame to all previous neurons.

The reasons behind choosing SMO classifier [4] are:

- Finding a Lagrange multiplier α 1 that violates the Karush–Ku hn–Tucker (KKT) conditions for the optimization problem.
- Picking a second multiplier α 2 and optimize the pair (α 1, α 2).
- Repeating steps 1 and 2 until convergence.

Also, the reasons for choosing Kstar classifier [5] are:

- Kstar operates on-the-fly, which means that it does not require the graph to be explicitly available and stored in the main memory. Portions of the graph will be generated as needed.
- Kstar can be guided using heuristic functions.

In the next section we will expose our new way to predict the new rank by using supervised learning [2].

4. Methods

From a network viewpoint, papers can be seen as nodes in a network and the citations between papers as edges (see Figs. 1 and 2).

As we see in Fig. 3, in the network, each paper node X links to another paper node Y through citation between themes. This network can help us to have more information about authors, papers, type of papers, etc. Then, like all transfer a model, the score is calculated by the count of the number of citations will be transferred to the referenced papers. Also, we must split our data into subsections to rank each paper into its division. As a case in point, we will treat information about different research publications in the field of computer science exclusively for Geographic Information System.

In this work, we take into account the following point:

- The papers with high number of citations reflect the importance and the prestige of the author.
- Scientific Gem [25] is always in the first rank in spite of their date of publication as a result of their recent citations.
- Recent publication has always less citation despite their newest contribution.

In any machine, the learning algorithm determines the good features which can provide very good results. The application of our algorithm [26] for ranking depends on:

- Paper Posted Time: The number of years since it has been published, by the formula:
  - A = Current Year-Year of Publication.
- Conference Score: The quality of any conference can be explored by the age of the conference, the continuity for the conference, the number of papers in the proceeding and the Digital Library involved.

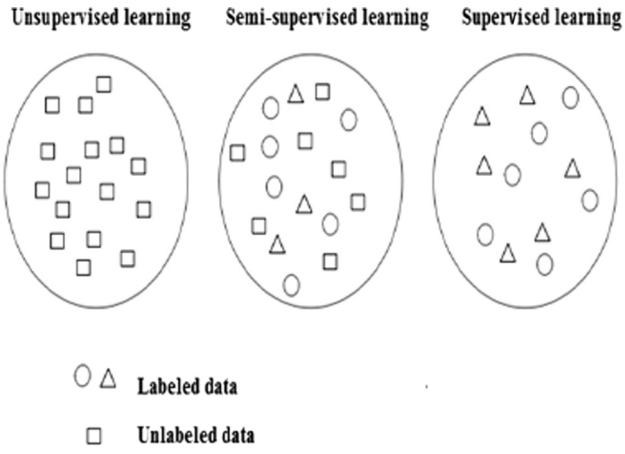

Fig. 1. Selection of data according to the category of learning.

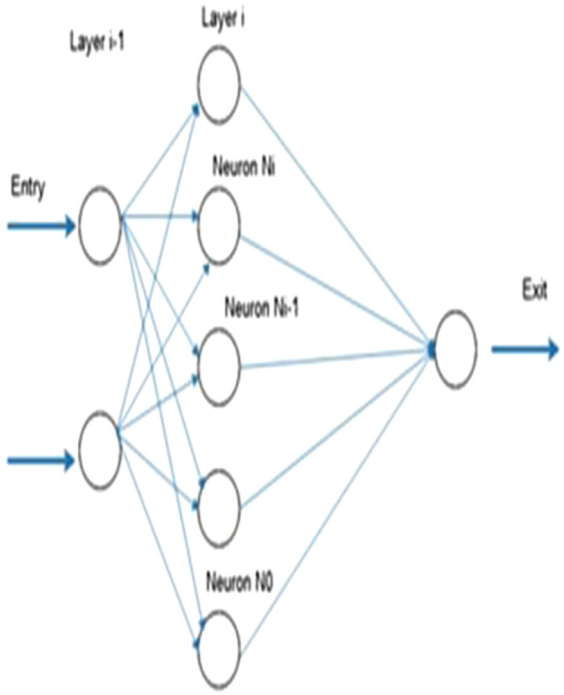

Fig. 2. Example of a neural network.

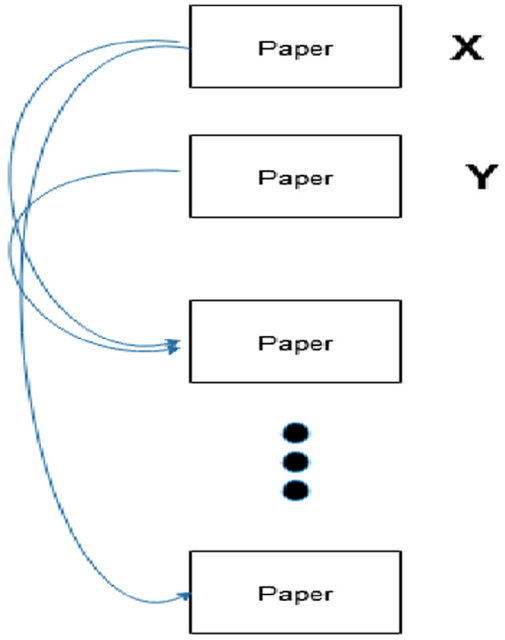

Fig. 3. The scientific papers network.

- Author Score: The number of authors of each paper, and the number of publications for each author is used to calculate author score, which is defined by:

$$\sum_{i=1}^{n} A_i DL\left(\frac{1}{NH}\right)$$

- Download Rates: The number of downloads from the official website of the journal reflects the importance for the given work in the paper.
- Keywords: The order of the keywords reflects the topics and the interest for the work.
- Publication Type: In general, the papers that are published in journals are more influential than the other types of publication venues, and the importance of conference is less influential than journals and higher than workshops.
- The Average Publication/Keyword:

This feature is given as below:

- 0: keyword not in the title of paper.
- 1: keyword in abstract.
- 2: keyword in the title of paper.
- 3: keyword in both title and abstract

The rank can be computed by the giving equation:

$$Rank = \sum_{i=1}^{n} A_i.DL\left(\frac{1}{NH}\right) + 0.2(A_p + A_{nbr}) + 0.3(\Delta\omega + \text{type}) \pm \text{PR(i)} \times \text{DL}\left(\frac{1}{1+\log A}\right)$$

- A_i: The number of papers published by the author.
- N: The total number of all authors for the paper.
- H: Constant with value 10.
- A_v: The average publication/keyword.
- A_{nbr}: The order of paper.
- $\Delta\omega$: Download rates.
- Type: The type of paper.
- PR(i): The score calculated by the PageRank algorithm.
- Log (A): Used to reduce the impact of old paper having the highest number of citations that are called "Scientific gem" [25].

In this paper, we examine the possibility of predicting the new rank for scientific papers to help researchers to find papers that they are looking for; we choose to work with Weka [27], as a collection of machine learning algorithms.

5. Results and discussion

5.1. Dataset

We give a great importance to the construction of training data and test data because of their influence on our experience in this work; for this reason, the mobile window strategy has been adopted.

The appropriate size of test data must respect certain conditions. First of all, it must not be too small because it will converge easily; this will have consequences on the accuracy of the prediction because, on the one hand, it has not given enough data to support the reasons enough and, on the other hand, it must not be too big because it is not necessary to converge (see Figs. 4–6).

We have chosen the moving windows strategy [28] compared to the sliding window strategy [29] because the prediction depends on the time factor, which is the case in our study. After the realization of some preliminary experience, we try to choose the two data model (test data and training data) which guarantees the lowest possible error rate.

In this research, we have exploited the bibliographic datasets from Thomson-Reuters Web of Science [30], which has information about different research publications in distinct fields of science, basic metadata for 1.935 SRPs in Research Areas of Science, and cover publications from 1996 to 2015. We use data from 1996 to 2015. It is to be noted that the Thomson-Reuters dataset that we used contained citation information till the year 2015 only.

5.2. Data pre-processing

The Web of Science dataset also contains information that is not useful in our algorithm. We need to pre-process the dataset to extract only the information that we will use in our algorithm. In data pre-processing, after the extraction and preparation of data, the final database contains: the title of paper, its author(s), keywords, paper posted time, the conference of publication, the given research paper cites, the download rates, also the average publication/keyword, article number and finally the order of paper.

For our experiment, we split our data in two sections: the first is the data before 2012 that represent our training dataset that permits us to compute the rank for the papers; the second section contains the data that we wish to predict.

From the point of view of the users who were searching in 2012 for scientific research papers, the only available data is the training dataset, which means papers before 2012. Our system can predict for our user the papers that he/she wants with the highest rank in his/her subject.

These experiments have been designed to find the pertinent metrics such as: paper-id, author score, and number of papers published, average download rates, number of citations. These metrics can be calculated from our training dataset.

In our application, we apply some rules:

- We don't eliminate not integral data.
- We don't rank the papers whose author doesn't remain in the training dataset.
- We calculate the average of the rank that obtains all papers for each first author.

5.3. Mathematical model

Our Mathematical model can be modeled in a set theory. The system is represented as follows:

S= {I, P, O}

Where

S = represent the system,

Where I = inputs, represent the given features,

P = {P_1,P_2...P_n}

Where P = Processes

P_1P_1 = check data in local server, P2 = store data at database,

O = {O_1, O_2, ... O_n}

Where O = outputs, represent the new rank for our papers.

We took the papers related to the topics of Geographic Information System in our data set by our algorithm [26] as the training dataset.

Rank values of our proposed algorithm are based on three principle points and depicted in table mentioned as supplementary material.

Fig. 4. The training dataset originating before 2012, and the evaluation dataset originating after 2012.

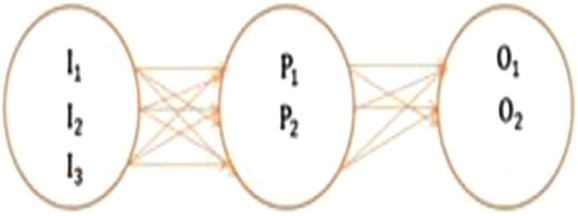

Fig. 5. Venn diagram of the proposed system.

- **Author and title** where the names of the authors and the titles of papers are presented.
- **DOI (Digital Object Identifier)** is a standardized method for the permanent identification of a published electronic object, a kind of permanent code of scientific articles. Each paper has its own DOI.
- **Journal and year** mean the date of paper publication and the journal where is submitted.

Now, our aim is to predict scientific research papers shown in Fig. 9, that are in the evaluation dataset whose authors are all in the training dataset and we calculate their futures as:

We will start by defining some parameters for our analysis:

- **Precision:** is the fraction of retrieved instances that are relevant, named Positive Predictive Value.
- **Recall:** is the fraction of relevant instances that are retrieved, named Sensitivity.
- **F-Measure:** is the harmonic mean of precision and recall; in other terms, it is the measure that combines precision and recall, and it is defined as:

$$F\text{-}Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

- **Cumulative Gain:** is the sum of the graded relevance values of all results in a search result list, where rel_i is the graded relevance of the result at position I; it is defined as:

$$CG_p = \sum_{i=1}^{p} rel_i$$

- **Discounted Cumulative Gain** [31]: is a particular rank position p; it is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)}$$

- **Normalized Discounted Cumulative Gain** [32]: search result lists vary in length depending on the query; it is defined as:

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

- **Average Precision:** summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, where P_n and R_n are the precision and recall at the nth threshold:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

- **Mean Average Precision** [33]: is the average of the precision value obtained for the set of k documents after the document is retrieved.

Q is $\{d_1 \ldots d_{mj}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then

$$MAP_d = \frac{1}{d} \sum_{j=1}^{d} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk})$$

```
@relation New-RANK

@attribute Paper-Id
@attribute Title
@attribute Author-score
@attribute Number-paper-published
@attribute Average-Download-rates
@attribute Average-number-of-citation
@attribute Rank{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29}

@data

1,"Participatory GIS to inform coral reef ecosystem management: Mapping human coastal and ocean uses in Hawaii",0.2,10,2389,1,?
2,"Public Participatory GIS",0.1,6,7458,0,?
3,"Modeling Flood Hazard Zones at the Sub-District Level with the Rational Model Integrated with GIS and Remote Sensing Approaches",0.3,7,3649,0,?
4,"Empirical PPGIS/PGIS mapping of ecosystem services: A review and evaluation",0.4,12,2459,4,?
5,"Using participatory GIS to examine social perception towards proposed wind energy landscapes",0.1,9,7561,0,?
6,"USE AND CONSEQUENCES OF PARTICIPATORY GIS IN A MEXICAN MUNICIPALITY: APPLYING A MULTILEVEL FRAMEWORK",0.3,7,4567,0,?
7,"On the criticality of mapping practices: Geodesign as critical GIS?",0.1,5,8560,0,?
8,"A participatory approach in GIS data collection for flood risk management, Muzarabani district, Zimbabwe",0.1,9,7956.1,?
9,"Shifting Boundaries of Volunteered Geographic Information Systems and Modalities: Learning from PGIS",0.099,6,1456,0,?
10,"Public participation GIS and neighbourhood recovery: using community mapping for economic development",0.05,6,4368,0,?
11,"A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio",0.1,10,6319,1,?
12,"The Digital Divide in Citizen-Initiated Government Contacts: A GIS Approach",0.3,5,2996,0,?
13,"GIS-Based Planning and Modeling for Renewable Energy: Challenges and Future Research Avenues",0.8,9,6795,3,?
14,"A participatory GIS approach to identify critical land degradation areas and prioritize soil conservation for mountainous olive groves (case study)",0.1,3,2978,0,?
15,"An empirical evaluation of workshop versus survey PPGIS methods",0.0852,6,6891,6,?
16,"GIS-Based Roughness Derivation for Flood Simulations: A Comparison of Orthophotos, LiDAR and Crowdsourced Geodata",0.3,8,9315,2,?
17,"Opportunities and Impediments for Open GIS",0.1,5,3189.6,?
18,"Which 'public'? Sampling effects in public participation GIS (PPGIS) and volunteered geographic information (VGI) systems for public lands management",0.0375,12,6975,13,?
19,"Discovering Spread Mode of Public Opinions in Incidents and Mapping it with GIS: a Case on Big Geospatial Data Analytics",0.15,7,3915,0,?
20,"Using participatory GIS to measure physical activity and urban park benefits",0.0375,13,9468,10,?
21,"Participatory GIS for strengthening transboundary marine governance in SIDS",0.1,6,3986,2,?
22,"The Global Landscape of GIS in Secondary Education",0.3,7,3956,2,?
23,"GIS and agent-based models for humanitarian assistance",0.2,9,8156,3,?
24,"The Spatial Politics of Affect and Emotion in Participatory GIS",0.2,6,4896,9,?
25,"A place-based approach to conservation management using public participation GIS (PPGIS)",0.025,9,6891,3,?
26,"A Participatory GIS Solution for Watershed Rehabilitation Project Management in the Changjiang and Pearl River Basins",0.3,2,1978,0,?
27,"Using public participation GIS (PPGIS) on the Geoweb to monitor tourism development preferences",0.025,8,6138,5,?
28,"PARTICIPATORY GIS FOR WATER PROVISION AND COMMUNITY PLANNING - CASE STUDY KOFFIEIKRAAL, SOUTH AFRICA",0.2,2,1625,0,?
29,"PARTICIPATORY GIS: EXPERIMENTATIONS FOR A 3D SOCIAL VIRTUAL GLOBE",0.3,1,540,0,?
```

Fig. 6. Data for prediction.

- **Geometric Mean Average Precision** [34]: is the geometric mean of the average precision values for an information retrieval system over a set of n query topics; it is defined as:

$$GMAP = \sqrt[n]{\prod_{n} AP_n}$$

The test set gives us the following results for the three classifiers after making some modification:

In Table 1, we can clearly see that the Multilayer Perceptron classifier has the highest $NDCG_{26}$ [32] compared to the tow classifiers and this can provide a good performance for prediction.

Our network for predicting new rank after using Multilayer Perceptron is shown in the figure below:

As we see in Fig. 7, our network contains in the input layer metrics: paper-id, author-score, number–paper-published, average-download-rates, average-number-citation, and one hidden layer with 17 nodes as the average between the number of inputs and outputs. Each connection node named neuron has a weight calculated from their inputs with a sigmoid function [35] as:

$$W_{next} = W + \Delta W$$

$$\Delta W = -learning\_rate \times gradient + momentum \times \Delta W_{previous}$$

Finally, we have the output layer with 29 classes representing the rank of scientific research papers. The predictions give us the result shown in Table 2.

5.4. Further comparison of the proposed algorithm

Now, we are analyzing the metrics used to predict our new rank. We propose four variant APC, APD, ADC and PDC explained as follow:

- New Rank (APC): a proposed variant wherein paper-id, author score, number–paper-published and average-number-citation are the parameters used.
- New Rank (APD): a proposed variant based on the same parameters of APC, but instead using average-number-citation we use average download-rates.
- New Rank (ADC): a proposed variant wherein we use paper-id, author score, average download-rates and average-number-citation.
- New Rank (PDC): a proposed variant based on paper-id, number-paper-published, average-download-rates and average-number-citation.

Table 1
Performance for the three classifiers.

	Multilayer Perceptron	SMO	Kstar
DCG_{26}	129.302	122.092	125.707
$IDCG_{26}$	162.678	162.677	162.683
$NDCG_{26}$	0.794	0.750	0.7727

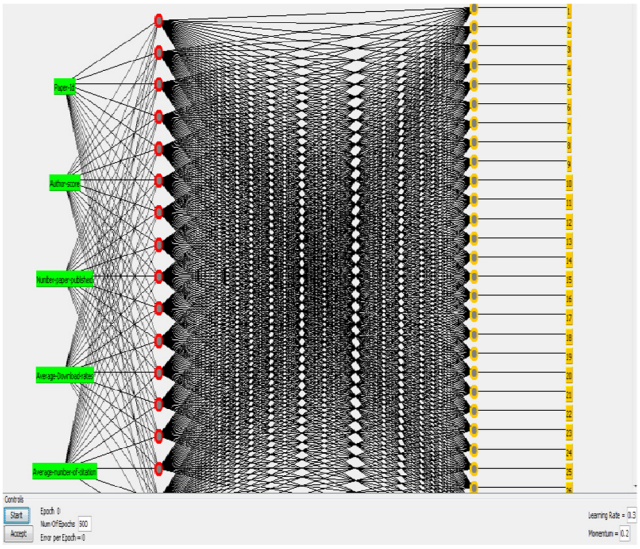

Fig. 7. The network for prediction.

In the figure below, we summarize the results for the four variants of our proposed new rank.

In order to define the suitable variant of our new rank the three parameters: Precision, Recall and F-measures should have higher values. As we see in the figure, we conclude that the APD variant satisfied this condition (see Fig. 8).

As we see in the previous sub-section, MAP is just an average precision which most frequently used in research papers. In contrast to MAP which can be considered as an arithmetic mean, GMAP is a geometric per precision; it used to highlight improvement for low performing subjects.

Table 2
Scientific research papers predicted by our new rank algorithm.

Rank	Author and Title	DOI	Journal, Year
1	Wilson: On the criticality of mapping practices: Geodesign as critical GIS?	https://doi.org//10.1016/j.landurbplan.2013.12.017	LANDSCAPE AND URBAN PLANNING, 2015
2	Brown et al.: An empirical evaluation of workshop versus survey PPGIS methods	https://doi.org//10.1016/j.apgeog.2014.01.008	APPLIED GEOGRAPHY, 2014
3	Mukherjee: Public Participatory GIS	https://doi.org//10.1111/gec3.12223	Geography Compass, 2015
4	Brown and Weber: A place-based approach to conservation management using public participation GIS (PPGIS)	https://doi.org//10.1080/09640568.2012.685628	JOURNAL OF ENVIRONMENTAL PLANNING AND MANAGEMENT, 2013
5	Brown and Weber: Using public participation GIS (PPGIS) on the Geoweb to monitor tourism development preferences	https://doi.org//10.1080/09669582.2012.693501	JOURNAL OF SUSTAINABLE TOURISM, 2013
6	Al-Wadaey and Ziadat: A participatory GIS approach to identify critical land degradation areas and prioritize soil conservation for mountainous olive groves (case study)	https://doi.org//10.1007/s11629-013-2827-x	JOURNAL OF MOUNTAIN SCIENCE, 2014
7	Mekonnen and Gorsevski: A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio	https://doi.org//10.1016/j.rser.2014.08.030	RENEWABLE & SUSTAINABLE ENERGY REVIEWS, 2015
8	Young and Gilmore: The Spatial Politics of Affect and Emotion in Participatory GIS	https://doi.org//10.1080/00045608.2012.707596	ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS, 2013
9	Thompson: Public participation GIS and neighbourhood recovery: using community mapping for economic development	https://doi.org//10.1504/IJDMMM.2015.067632	INTERNATIONAL JOURNAL OF DATA MINING MODELLING AND MANAGEMENT, 2015
10	Baldwin et al.: Participatory GIS for strengthening trans boundary marine governance in SIDS	https://doi.org//10.1111/1477-8947.12029	NATURAL RESOURCES FORUM, 2013
11	Pozzebon et al.: Use and consequences of participatory GIS in Mexican municipality: applying a multilevel framework	https://doi.org//10.1590/S0034-759020150305	RAE-REVISTA D'ADMINISTRACAO D'EMPRESAS, 2015
12	Zhang et al.: Discovering Spread Mode of Public Opinions in Incidents and Mapping it with GIS: a Case on Big Geospatial Data Analytics	https://doi.org//10.1109/Agro-Geoinformatics.2014.6910597	AGRO-GEOINFORMATICS, 2014
13	Sui: Opportunities and Impediments for Open GIS	https://doi.org//10.1111/tgis.12075	TRANSACTIONS IN GIS, 2014
14	Asare-Kyei et al.: Modeling Flood Hazard Zones at the Sub-District Level with the Rational Model Integrated with GIS and Remote Sensing Approaches	https://doi.org//10.3390/w7073531	WATER, 2015
15	Kerski et al.: The Global Landscape of GIS in Secondary Education	https://doi.org//10.1080/00221341.2013.801506	JOURNAL OF GEOGRAPHY, 2013
17	Levine and Feinholz: Participatory GIS to inform coral reef ecosystem management: Mapping human coastal and ocean uses in Hawaii	https://doi.org//10.1016/j.apgeog.2014.12.004	APPLIED GEOGRAPHY, 2015
18	McCall et al.: Shifting Boundaries of Volunteered Geographic Information Systems and Modalities: Learning from PGIS	/1234	ACME: An International Journal for Critical Geographies, 2015
19	Brown and Fagerholm: Empirical PPGIS/PGIS mapping of ecosystem services: A review and evaluation	https://doi.org//10.1016/j.ecoser.2014.10.007	ECOSYSTEM SERVICES, 2015
20	Song et al.: A Participatory GIS Solution for Watershed Rehabilitation Project Management in the Changjiang and Pearl River Basins	https://doi.org//10.2991/rsete.2013.114	Advances in Intelligent Systems Research, 2013
21	Panek and Van Heerden: Participatory GIS for water provision and community planning - Case study Koffieikraal, South Africa	https://doi.org//10.5593/SGEM2013/BB2.V1/S11.030	Cartography and GIS, 2013
22	Brovelli et al.: Participatory GIS: Experimentations for a 3D social virtual globe	https://doi.org//10.5194/isprsarchives-XL-2-W2-13-2013	International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2013
23	Chingombe et al.: A participatory approach in GIS data collection for flood risk management, Muzarabani district, Zimbabwe	https://doi.org//10.1007/s12517-014-1265-6	ARABIAN JOURNAL OF GEOSCIENCES, 2015
24	Lombard: Using participatory GIS to examine social perception towards proposed wind energy landscapes	https://doi.org//10.17159/2413-3051/2015/v26i2a2195	JOURNAL OF ENERGY IN SOUTHERN AFRICA, 2015
25	Crooks and Wise: The role of Public Participatory Geographical Information Systems (PPGIS) in coastal decision-making processes: An example from Scotland, UK	https://doi.org//10.1016/j.compenvurbsys.2013.05.003	COMPUTERS ENVIRONMENT AND URBAN SYSTEMS, 2013
26	Dorn et al.: GIS-Based Roughness Derivation for Flood Simulations: A Comparison of Orthophotos, LiDAR and Crowdsourced Geodata	https://doi.org//10.3390/rs6021739	REMOTE SENSING, 2014
27	Brown et al.: Which 'public'? Sampling effects in public participation GIS (PPGIS) and volunteered geographic information (VGI) systems for public lands management	https://doi.org//10.1080/09640568.2012.741045	JOURNAL OF ENVIRONMENTAL PLANNING AND MANAGEMENT, 2014
28	Brown et al.: Using participatory GIS to measure physical activity and urban park benefits	https://doi.org//10.1016/j.landurbplan.2013.09.006	LANDSCAPE AND URBAN PLANNING, 2014
29	Resch et al.: GIS-Based Planning and Modeling for Renewable Energy: Challenges and Future Research Avenues	https://doi.org//10.3390/ijgi3020662	ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION, 2014

On the other hand, we should compare MAP and GMAP for the four variants. We clearly see in the Fig. 9 in term of APD, ADC and PDC the values are slightly close to each other compared to APC which has superior values.

In Fig. 10, we present a comparison between GMAP and MAP in term of the proposed new rank. We can clearly see that for each ranked paper GMAP and MAP are close to each other, instead of some cases wherein we find that the values of GMAP are very less than MAP.

MAP and GMAP may be seen as similar measures of average ranking effectiveness of a system.

To sum up, the Figs. 9 and 10 shows that GMAP values are less than MAP and this lead to a perform ranking and at the same time reducing errors in our proposed system.

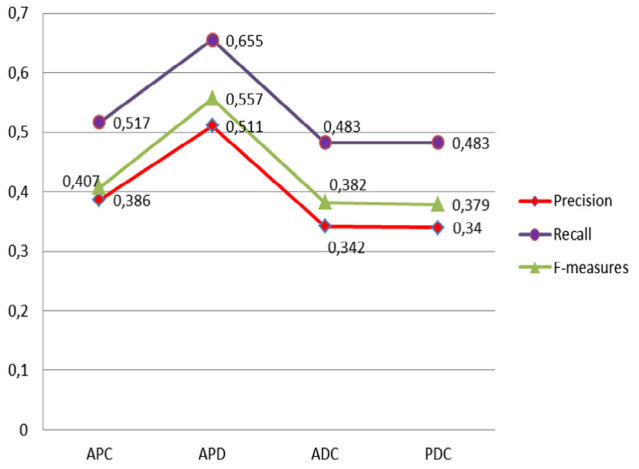

Fig. 8. Comparison of the performance.

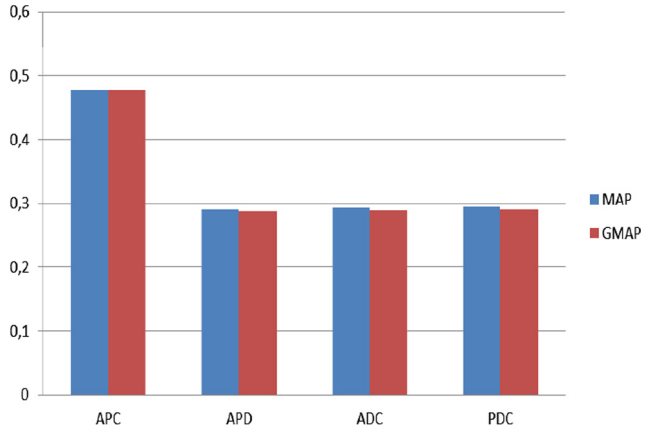

Fig. 9. MAP vs. GMAP in different variants.

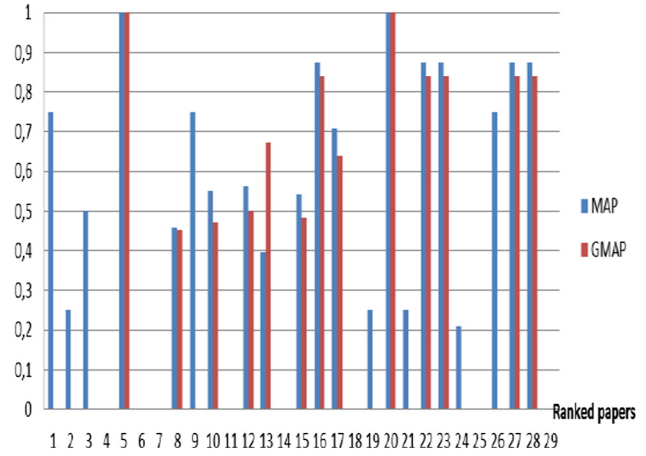

Fig. 10. GMAP vs. MAP in the proposed New Rank.

6. Conclusion

In this paper, we propose a new approach for predicting the new rank for scientific research papers. Our experimental evaluation has shown the efficient of the utilization of machine learning algorithm in the discipline of ranking. We provide an algorithm that use different metrics such us (including but not restricted to paper-id, author-score, number–paper-published, average-download-rates, average-number-citation) in a one network. Moreover, we provide a comparison of the metrics and ranked them conforming to their prediction ability using metrics analysis algorithm, we think that this work can help researcher in other discipline such us: financial sector, policies investigation, and terrorist behavior.

For future work, we plan to test our algorithm on additional datasets in order to determine how robust it is to the different values of parametrs, and in different datasets, also we plane to make a survey to appraise the results by users.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.aci.2018.02.002.

References

[1] N.W. Alkharouf, D.C. Jamison, B.F. Matthews, Online analytical processing (OLAP): a fast and effective data mining tool for gene expression databases, J. Biomed. Biotechnol. 2005 (2005) 181–188, https://doi.org/10.1155/JBB.2005.181.
[2] M. Darnstädt, H.U. Simon, B. Szörényi, Supervised learning and co-training, Algorithmic Learn. Theory 519 (2014) 68–87, https://doi.org/10.1016/j.tcs.2013.09.020.
[3] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, M. Ettaouil, Multilayer perceptron: architecture optimization and training, Int. J. Interact. Multimed. Artif. Intell. 4 (2016) 26, https://doi.org/10.9781/ijimai.2016.415.
[4] X. Shao, K. Wu, B. Liao, Single directional SMO algorithm for least squares support vector machines [WWW Document], Intell. Neurosci. Comput. (2013), https://doi.org/10.1155/2013/968438.
[5] S. Lee, M. Park, J. Park, H. Na, M. Kwon, Operator interface programs for KSTAR operation, Fusion Eng. Des. 88 (2013) 2835–2841, https://doi.org/10.1016/j.fusengdes.2013.05.008.
[6] K. Slim, A. Dupré, B. Le Roy, Impact factor: an assessment tool for journals or for scientists?, Anaesth Crit. Care Pain Med. 36 (2017) 347–348, https://doi.org/10.1016/j.accpm.2017.06.004.
[7] Y. Du, W. Liu, X. Lv, G. Peng, An improved focused crawler based on semantic similarity vector space model, Appl. Soft Comput. 36 (2015) 392–407, https://doi.org/10.1016/j.asoc.2015.07.026.
[8] F. Zhuang, G. Karypis, X. Ning, Q. He, Z. Shi, Multi-view learning via probabilistic latent semantic analysis, Inf. Sci. 199 (2012) 20–30, https://doi.org/10.1016/j.ins.2012.02.058.
[9] S. Robertson, H. Zaragoza, M. Taylor, Simple BM25 Extension to Multiple Weighted Fields, in: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04. ACM, New York, NY, USA, 2004, pp. 42–49. https://doi.org/10.1145/1031171.1031181.
[10] F. Lv, H. Zhang, J.g. Lou, S. Wang, D. Zhang, J. Zhao, CodeHow: Effective Code Search Based on API Understanding and Extended Boolean Model (E), in: 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE). Presented at the 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2015, pp. 260–270. https://doi.org/10.1109/ASE.2015.42.
[11] X. Liu, H. Lin, C. Zhang, An improved HITS algorithm based on page-query similarity and page popularity, J. Comput. 7 (2012), https://doi.org/10.4304/jcp.7.1.130-134.
[12] X. Tan, A new extrapolation method for PageRank computations, J. Comput. Appl. Math. 313 (2017) 383–392, https://doi.org/10.1016/j.cam.2016.08.034.
[13] A. Bougouin, F. Boudin, B. Daille, Topicrank: Graph-based topic ranking for keyphrase extraction, in: International Joint Conference on Natural Language Processing (IJCNLP), 2013. pp. 543–551.
[14] P. Jomsri, S. Sanguansintukul, W. Choochaiwattana, CiteRank: combination similarity and static ranking with research paper searching, Int. J. Internet Technol. Secur. Trans. 3 (2011) 161–177, https://doi.org/10.1504/IJITST.2011.039776.
[15] M.A. Hasson, S.F. Lu, B.A. Hassoon, Scientific research paper ranking algorithm PTRA: a tradeoff between time and citation network, Appl. Mech. Mater. 551 (2014) 603–611, https://doi.org/10.4028/www.scientific.net/AMM.551.603.
[16] F. Stahl, I. Jordanov, An overview of the use of neural networks for data mining tasks, Wiley Interdiscip Rev. Data Min. Knowl. Discov. 2 (2012) 193–208, https://doi.org/10.1002/widm.1052.
[17] Y. Zheng, B. Jeon, L. Sun, J. Zhang, H. Zhang, Student's t-hidden markov model for unsupervised learning using localized feature selection, IEEE Trans. Circuits Syst. Video Technol. 1–1 (2017), https://doi.org/10.1109/TCSVT.2017.2724940.
[18] F.-J. González-Serrano, Á. Navia-Vázquez, A. Amor-Martín, Training support vector machines with privacy-protected data, Pattern Recognit. 72 (2017) 93–107, https://doi.org/10.1016/j.patcog.2017.06.01.

[19] W.W.Y. Ng, X. Zhou, X. Tian, X. Wang, D.S. Yeung, Bagging–boosting-based semi-supervised multi-hashing with query-adaptive re-ranking, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.042.
[20] J. Yao, Q. Mao, S. Goodison, V. Mai, Y. Sun, Feature selection for unsupervised learning through local learning, Pattern Recognit. Lett. 53 (2015) 100–107, https://doi.org/10.1016/j.patrec.2014.11.006.
[21] K.M. Kumar, A.R.M. Reddy, An efficient k-means clustering filtering algorithm using density based initial cluster centers, Inf. Sci. 418–419 (2017) 286–301, https://doi.org/10.1016/j.ins.2017.07.036.
[22] Z. Zainuddin, O. Pauline, An effective fuzzy C-means algorithm based on symmetry similarity approach, Appl. Soft Comput. 35 (2015) 433–448, https://doi.org/10.1016/j.asoc.2015.06.021.
[23] X. Zhu, Semi-Supervised Learning, in: Encyclopedia of Machine Learning, Springer, Boston, MA, 2011, pp. 892–897, https://doi.org/10.1007/978-0-387-30164-8_749.
[24] S. Chen, S. Zhu, Y. Yan, Robust visual tracking via online semi-supervised co-boosting, Multimed. Syst. 22 (2016) 297–313, https://doi.org/10.1007/s00530-015-0459-4.
[25] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with Google's PageRank algorithm, J. Informetr. 1 (2007) 8–15, https://doi.org/10.1016/j.joi.2006.06.001.
[26] M. El Mohadab, B. Bouikhlaene, S. Safi, Towards an efficient algorithm for ranking scientific research papers, in: 2017 2nd IEEE international scientific event on internet of things: Recent innovations and challenges (SEIT). Presented at the 2017 2nd IEEE international scientific event on internet of things: Recent innovations and challenges (SEIT), Rabat, Morocco.
[27] S. Lievens, B. De Baets, Supervised ranking in the weka environment, Inf. Sci. 180 (2010) 4763–4771, https://doi.org/10.1016/j.ins.2010.06.014.
[28] J. Valluru, P. Lakhmani, S.C. Patwardhan, L.T. Biegler, Development of moving window state and parameter estimators under maximum likelihood and Bayesian frameworks, DYCOPS-CAB 2016 (60) (2017) 48–67, https://doi.org/10.1016/j.jprocont.2017.08.007.
[29] I. Chakroun, T. Haber, T.J. Ashby, SW-SGD: The Sliding Window Stochastic Gradient Descent Algorithm. Procedia Comput. Sci., International Conference on Computational Science, ICCS 2017, 12–14 June 2017, Zurich, Switzerland 108, 2017, 2318–2322. https://doi.org/10.1016/j.procs.2017.05.082.
[30] Web of Science library. Available at: http://www.webofknowledge.com [accessed in 2016].
[31] G. Dupret, B. Piwowarski, Model Based Comparison of Discounted Cumulative Gain and Average Precision. Sel. Pap. 18th Int. Symp. String Process. Inf. Retr. SPIRE 2011 18, 49–62. https://doi.org/10.1016/j.jda.2012.10.002.
[32] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, W. Chen. A theoretical analysis of NDCG type ranking measures. ArXiv13046480 Cs Stat, 2013.
[33] M. Thelwall, The precision of the arithmetic mean, geometric mean and percentiles for citation data: an experimental simulation modelling approach, J. Informetr. 10 (2016) 110–123, https://doi.org/10.1016/j.joi.2015.12.001.
[34] S. Robertson, On GMAP: and other transformations. (2006), https://doi.org/10.1145/1183614.1183630.
[35] A. Iliev, N. Kyurkchiev, S. Markov, On the approximation of the step function by some sigmoid functions, Math. Comput. Simul., Biomath 2014 and Biomath 2015 133 (2017) 223–234, https://doi.org/10.1016/j.matcom.2015.11.